RDMA通信方式的介绍

笔记本:	内存计算		
创建时间:	2018/11/21 17:05	**更新时间:**	2018/11/22 10:57
作者:	simba_wei		

RDMA：remote direct memory access方式的简称，即绕开kernel和协议的内存直接获取方式。
RDMA的通讯方式通常可以分为两种同时通讯方式

- 单边（one-sided）读写（read/write）操作
- 双边（two-sided）接受发送（send/receive）操作

这些RDMA的操作有如下的几个特点：

1. 应用执行RDMA操作时是应用**直接向NIC（网卡）上发送请求**，不需要内存拷贝且绕开CPU
2. **NIC读取缓冲内容**，然后直接传送到远程的NIC
3. 在网络上传输的RDMA 信息**包含目标虚拟地址、内存钥匙和数据本身。请求既可以完全在用户空间中处理(通过轮询用户级完成排列)，又或者在应用一直睡眠到请求完成时的情况下通过系统中断处理。**RDMA 操作使应用可以从一个远程应用的内存中读数据或向这个内存写数据。
4. 目标NIC 确认内存钥匙，**直接将数据写人应用缓存中。**用于操作的远程虚拟内存地址包含在RDMA信息中。

无论是单边还是双边操作，通信双方在远端都需要建立一条I/O channel。这里需要注意的是，每一个应用都对应着一条独立的I/O channel。每一个I/O channel都对应着一组QP （queue pair）分别放置在通信双方两端。每一对QP都由send queue（SQ）和receive queue（RQ）构成，这里我们提到的SQ和RQ都可以统称为WQ（work queue）。这些队列中管理着各种类型的消息。每一个应用的QP都会映射到自己应用所在的虚拟地址空间（这就意味着：在RNIC卡上的queue直接能够访问着应用对应的内存）。除了QP的两种队列之外，RDMA还会提供一种CQ（complete queue）。CQ用来知会应用程序SQ或RQ上的任务已经被处理完成了。

下面介绍一下单边的读写操作：
One-sided RDMA Read：A节点读取B节点数据的流程

1. 首先A、B两个节点间建立连接，即创建QP。QP创建好了之后就意味着A和B都已经知晓了对方的WQ
2. A要读取的数据存储在B的buffer中，具体地址为VB。B将VB的地址以及对应的这块地址的操作权限发送到A节点。
3. 在发送完消息后，B节点会往自己的WQ中注册一个WR，用于接收A的返回状态。
4. A在收到请求后会把自己用来存储数据的地址封装到RDMA的read请求中，发送到对应的B节点。B节点的WR任务收到A的读请求后，两节点开始通讯。
5. 数据传输完成后，A会告诉B的CQ，传输任务已经完成了。

One-sided RDMA Write：A节点和B节点写数据的流程：

1. 首先A、B两个节点间建立连接，即创建QP。QP创建好了之后就意味着A和B都已经知晓了对方的WQ
2. 数据remote目标存储buffer地址VB，注意VB应该提前注册到B的RNIC(并且它是一个Memory Region)，并拿到返回的local key，相当于RDMA操作这块buffer的权限。
3. B把数据地址VB，key封装到专用的报文传送到A，这相当于B把数据buffer的操作权交给了A。同时B在它的WQ中注册进一个WR，以用于接收数据传输的A返回的状态。

4. A在收到B的送过来的数据VB和R_key后，RNIC会把它们连同自身发送地址VA到封装RDMA WRITE请求，这个过程A、B两端不需要任何软件参与，就可以将A的数据发送到B的VB虚拟地址。
5. A在发送数据完成后，会向B返回整个数据传输的状态信息。
6. 单边操作传输方式是RDMA与传统网络传输的最大不同，只需提供直接访问远程的虚拟地址，无须远程应用的参与其中，这种方式适用于批量数据传输。

Two-sided RDMA Send/Receive：A节点和B节点的双边通信操作：
RDMA中SEND/RECEIVE是双边操作，即必须要远端的应用感知参与才能完成收发。在实际中，SEND/RECEIVE多用于连接控制类报文，而数据报文多是通过READ/WRITE来完成的。

1. 首先，A和B都要创建并初始化好各自的QP，CQ
2. A和B分别向自己的WQ中注册WQE，对于A，WQ=SQ，WQE描述指向一个等到被发送的数据；对于B，WQ=RQ，WQE描述指向一块用于存储数据的Buffer。
3. A的RNIC异步调度轮到A的WQE，解析到这是一个SEND消息，从Buffer中直接向B发出数据。数据流到达B的RNIC后，B的WQE被消耗，并把数据直接存储到WQE指向的存储位置。
4. AB通信完成后，A的CQ中会产生一个完成消息CQE表示发送完成。与此同时，B的CQ中也会产生一个完成消息表示接收完成。每个WQ中WQE的处理完成都会产生一个CQE。
5. 双边操作与传统网络的底层Buffer Pool类似，收发双方的参与过程并无差别，区别在零拷贝、Kernel Bypass，实际上对于RDMA，这是一种复杂的消息传输模式，多用于传输短的控制消息。